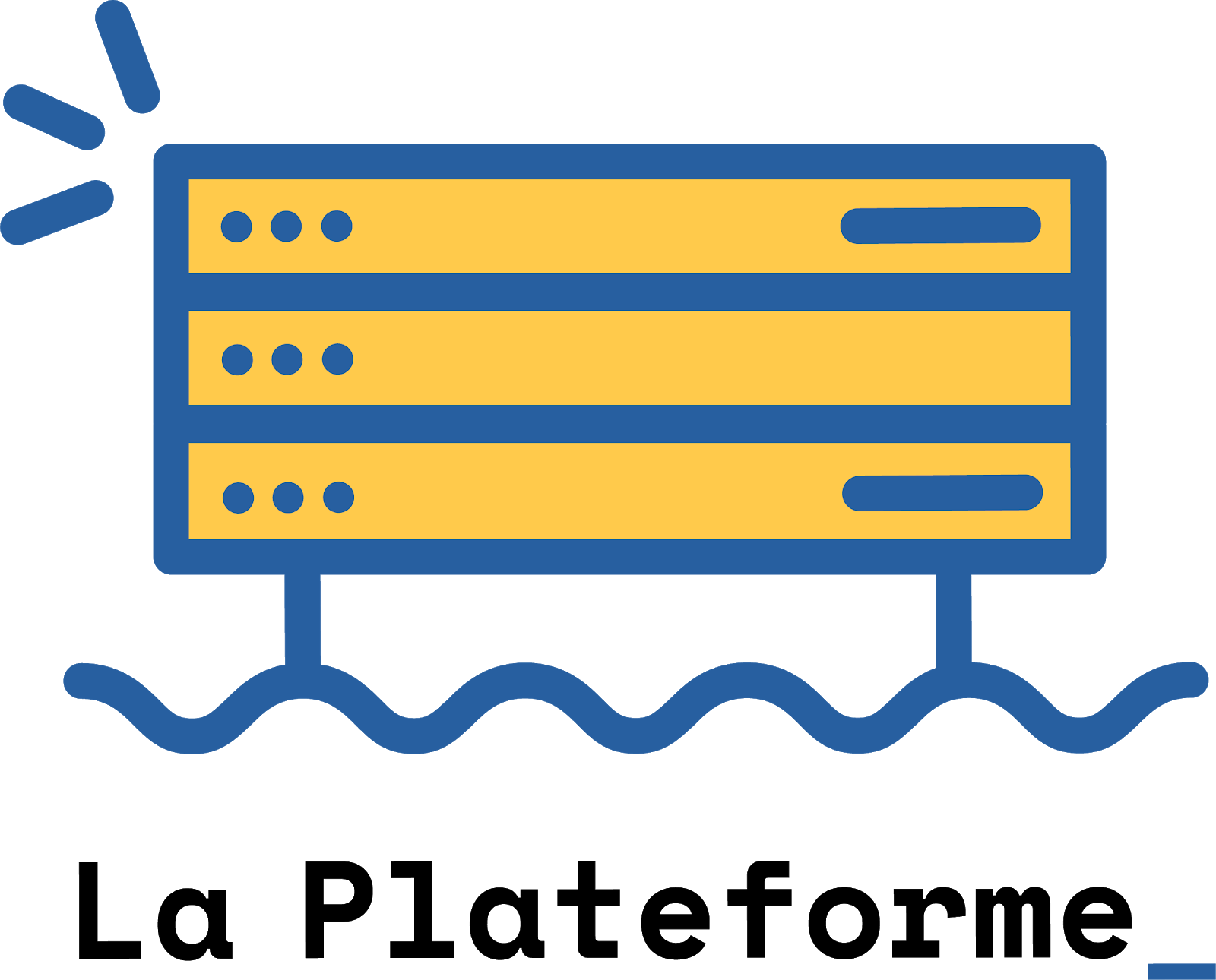
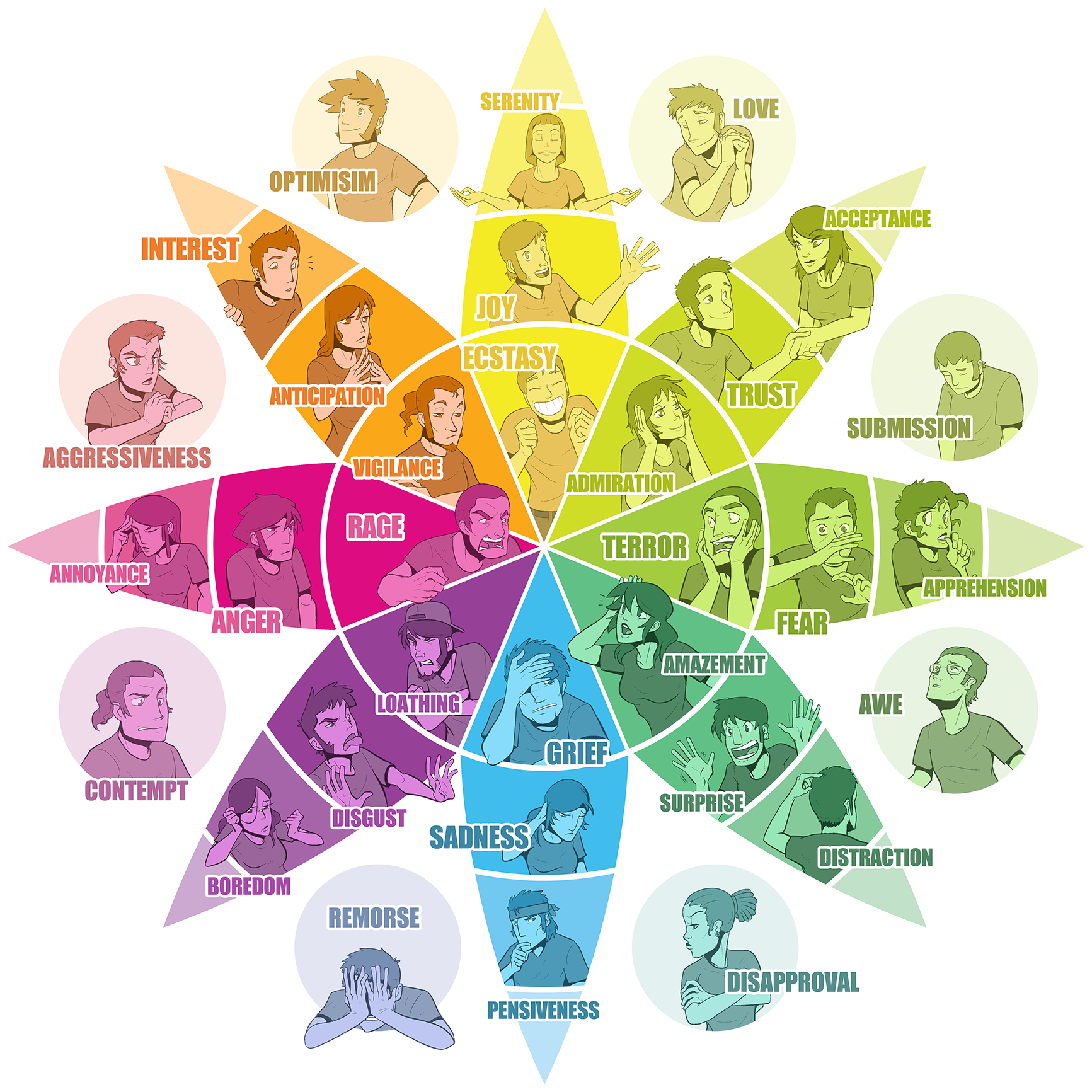
****

**La Plateforme\_**

**La roue des émotions**



****



Construit d’après les travaux du psychologue américain Robert Plutchik, la roue des émotions est un modèle des émotions humaines et peut facilement servir à définir des personnages, ainsi que leur évolution dans une trame narrative. Est-il possible d'identifier des émotions dans des phrases narratives issues de communications écrites ?

## **Contexte du projet**

Depuis quelques années, les dispositifs de communication médiatisée par ordinateur (CMO) sont massivement utilisés, aussi bien dans les activités professionnelles que personnelles. Ces dispositifs permettent à des participants distants physiquement de communiquer. La plupart implique une communication écrite médiatisée par ordinateur (CEMO) : forums de discussion, courrier électronique, messagerie instantanée. Les participants ne s’entendent pas et ne se voient pas mais peuvent communiquer par l’envoi de messages écrits, qui combinent, généralement, certaines caractéristiques des registres écrit et oral (Marcoccia, 2000a ; Marcoccia, Gauducheau, 2007 ; Riva, 2001).

Imaginez que vous souhaitez savoir ce qui se passe derrière votre écran d'ordinateur, quels sont vos contacts les plus actifs et quelle est leur personnalité (pas banal comme question !!). Vous allez alors vous lancer dans l’analyse de leur narration et tenter d’extraire quelle émotion se dégage de chacune des phrases.

Dans de nombreuses organisations, nous sommes amenés à utiliser des outils de discussion textuels et nous construisons nos relations sociales et professionnelles autour de ces dispositifs. Pour entretenir des rapports sociaux stables, sereins, de confiance et efficaces, au travers des outils de communication écrites, lorsqu'il n'est pas possible d'avoir la visio (avec caméra), il est nécessaire de détecter des éléments "clés" dans les channels de discussions / mails qui nous permettront de déceler de la colère, de la frustration, de la tristesse ou encore de la joie de la part d'un collègue ou d'un amis pour adapter nos relations sociales. En tant qu'expert en data science, nous allons vous demander de développer un modèle de machine learning permettant de classer les phrases suivant l'émotion principale qui en ressort.

Pour des questions d’ordre privé, nous ne vous demanderons pas de nous communiquer les conversations provenant de votre réseau social favori ou de vos emails mais nous allons plutôt vous proposer deux jeux de données contenant des phrases, ces fichiers ayant déjà été annoté.\_

Vous devrez proposer plusieurs modèles de classification des émotions et proposer une analyse qualitative et quantitative de ces modèles en fonction de critères d'évaluation. Vous pourrez notamment vous appuyer sur les outils de reporting des librairies déjà étudiées. Vous devrez investiguer aux travers de librairies d'apprentissage automatique standards et de traitement automatique du langage naturel (par exemple NLTK) comment obtenir les meilleures performance de prédiction possible en prenant en compte l'aspect multi-class du problème et en explorant l'impact sur la prédiction de divers prétraitement tel que la suppression des stop-words, la lemmatisation et l'utilisation de n-grams, et différente approche pour la vectorisation.

**Vous devrez travailler dans un premier temps avec le jeu de données issu de Kaggle pour réaliser vos apprentissage et l'évaluation de vos modèles.**

**Dans l'objectif d'enrichir notre prédictions nous souhaitons augmenter notre jeu de données. Vous devrez donc travailler dans un deuxième temps avec le jeux de données fournie, issue de data.world afin de :**

1. **Comparez d'une part si les résultats de classification sur votre premier jeu de données sont similaires avec le second. Commenter.**
2. **Combiner les deux jeux donnés pour tenter d'améliorer vos résultats de prédiction.**
3. **Prédire les nouvelles émotions présentes dans ce jeu de données sur les messages du premier, et observer si les résultats sont pertinents.**

Votre travail doit permettre une visualisation dédiée à l'analyse et au traitement des données (histogramme, bubble chart, scatterplot etc), avec au minimum :

1. **Un histogramme représentant la fréquence d’apparition des mots (commenter)**
2. **Un histogramme des émotions (commenter)**

Il vous est demandé de comparer les résultats, d'au moins 4 classifiers, présentés dans un tableau permettant de visualiser vos mesures (notamment le Naïve Bayes, le SVM, SGD ainsi que le KNN). Réaliser une veille sur ces quatres classifieurs. Que pouvez-vous dire de leur fonctionnement ? Quelles sont les différences majeures entre ces 4 classifieurs.

Pour compléter cette analyse, on pourra visualiser des courbes de rappel, de précision (permette de tracer la précision et le rappel pour différents seuils de probabilités), un rapport de classification (un rapport de classification visuel qui affiche la precision, le recall, le f1-score, support, une matrice de confusion ou encore une graphique permettant de visualiser les mots les plus représentatif associé à chaque émotions.

Créer une application client/serveur (Flask) permettant à un utilisateur d'envoyer du texte via un champ de recherche (ou un fichier sur le disque du client) et de lui retourner l'émotion du texte envoyé.

**(Bonus)** la roue des émotions du document (exemple: quelle proportion de chacune des émotions contient le document ?)

**Ressources**

* <https://www.actuia.com/contribution/victorbigand/tutoriel-tal-pour-les-debutants-classification-de-texte/>
* <https://bbengfort.github.io/tutorials/2016/05/19/text-classification-nltk-sckit-learn.html>
* <https://www.nltk.org/>
* <https://medium.com/neuronio/from-sentiment-analysis-to-emotion-recognition-a-nlp-story-bcc9d6ff61ae>
* <https://realpython.com/sentiment-analysis-python/#how-classification-works>
* <https://www.kaherecode.com/tutorial/demarrer-avec-flask-un-micro-framework-python>
* <https://openclassrooms.com/fr/courses/4425066-concevez-un-site-avec-flask>

## **Livrables**

- Notebook résumant votre travail